SUMIKOLAH: JURNAL ILMU PENDIDIKAN

Volume 1, Nomor 2, Agustus 2023 (hal. 1-14) | e-ISSN 2985-3842

The Development of an English Achievement Test at a Private Senior High School in Manado

Amelia Cindy Mogi*[1], Alfrits Roul Sinadia [2]
[1, 2]Fakultas Keguruan dan Ilmu pendidikan, Universitas Klabat, Manado, Indonesia
e-mail:[1*]s21530007@student.unklab.ac.id, [2]alfritssinadia@unklab.ac.id

Abstract

This research was conducted due to the unavailability of teacher-made tests developed based on the principles of developing quality tests for use in schools. This study aims to analyze the quality of a Grade X English test items developed together with an English teacher at a private high school in Manado City, North Sulawesi. The developed test was intended to measure student achievement in the middle of the semester or called the midterm exam. The test consisted of 50 multiple-choice items that were tested on 133 Grade X students at the school. The results of the analysis show the quality of each item in terms of difficulty, discriminating power, and effectiveness of distractors. The final evaluation results on the quality of each item showed that, out of 50 items, there were only 11 items that met the category of qualified item difficulty level and differentiating power. Some items have one, two, or three distractors that do not function properly or are ineffective for their function as distractors. For this reason, all ineffective distractors must be corrected so that the items can be used to measure student achievement.

Keywords— test item analysis, test item quality, item difficulty, item discrimination power, distractor effectiveness

Abstrak

Penelitian ini dilakukan karena tidak tersedianya tes buatan guru yang dikembangkan berdasarkan prinsip-prinsip pengembangan tes yang berkualitas untuk digunakan di sekolah. Penelitian ini bertujuan untuk menganalisis kualitas butir soal mata pelajaran bahasa Inggris Kelas X yang dikembangkan bersama dengan guru bahasa Inggris di sebuah SMA swasta di Kota Manado, Sulawesi Utara. sekolah. Tes yang dikembangkan merupakan tes yang diperuntukkan untuk mengukur prestasi belajar siswa di tengah semester atau ujian tengah semester. Tes ini terdiri dari 50 butir pilihan ganda yang diujicobakan kepada 133 siswa Kelas X di sekolah tersebut. Hasil analisis kualitas butir soal menunjukkan kualitas setiap butir dalam hal tingkat kesukaran, daya pembeda, dan efektivitas distraktor. Hasil evaluasi akhir mengenai kualitas setiap butir menunjukkan bahwa dari 50 butir soal hanya terdapat 11 butir yang memenuhi kategori tingkat kesukaran dan daya pembeda yang baik. Beberapa butir soal memiliki satu, dua, atau tiga distraktor yang tidak berfungsi dengan baik atau tidak efektif untuk fungsinya sebagai pengecoh. Untuk itu, semua distraktor yang tidak efektif harus diperbaiki agar butir soal tersebut dapat digunakan untuk mengukur prestasi siswa.

Kata kunci— analisis butir soal, kualitas butir soal, tingkat kesukaran butir, daya pembeda butir, efektivitas distraktor

INTRODUCTION

One of the main tasks that teachers must do well in classrooms is evaluating students' achievement. Evaluation is needed in educational activities because "to reach the purpose of the instructional activities, teachers conduct evaluative activities to measure how far the students understand the material" (Munadliroh, 2015, p. 2). These activities have become such normality in education that some people even see educational evaluation as a tool for measuring the school's value. With the need for knowledge on the performance of profit schools, public schools, or even non-profit schools, evaluation has become one of the key sources of that information (Grayson, 2012). In other words, the performance of a school depends on the quality of the evaluation. Evaluation and assessment are also good ways for teachers to know if they are using the right strategies in teaching. Baranovskaya and Shaforostova (2017) stated that teachers gain valuable information in correcting their teaching strategies by using evaluation and assessment. When teachers understand what is the best method of teaching that subject, their performance in teaching will also increase. Those are some of the important things about having an evaluation in education.

Evaluations in schools are done by using instruments such as tests to measure students' achievement. Kubiszyn and Borich (2013) stated assessments are usually circled and related to the subjectivity of the teacher, but a teacher can use the results of tests to assess objectively. To prove that teachers are not biased or pressured, they use tests in evaluating their students. How good the test items determine the quality of the evaluation (Kusumawati & Hadi, 2018). If teachers can improve the quality of each test item, they will improve the quality of the tests (Reynolds, Livingston, & Wilson, 2009). Teachers cannot just create the test items sloppily without any knowledge of how to design the questions. The development of tests must be thought thoroughly by teachers because "designing a test, formulating items, and processing grade is a complete science" (Hussain & Sajid, 2015, p. 725). To put it simply, creating test items is knowledge that must be studied, developed, and mastered for years and does not happen overnight. Therefore, a good test is needed to measure students' real performance in class.

Though the skills of designing tests are important, some teachers unfortunately still lack those skills. That means teachers need to broaden their knowledge around evaluation and testing. Teachers disregard the importance of designing a good quality test because "some of the teachers see assessment mainly for the purpose of grading the pupils" (Opara & Magnus-Arewa, 2017, p. 48). They see tests only as a tool to grade the students because they have to grade the students, not because of the importance of evaluation. Next, after knowing the importance of designing a good test, teachers need to know if they are on the right track in various ways. The reason for this is that some other reasons can disrupt students' perceptions even when the test items are constructed using specific and structured guidelines (Karkal & Kundapur, 2016). One of the ways to know the effectiveness of tests is by using item analysis. In developing tests that measure students' knowledge of a certain subject, "item analysis plays an important role both in contributing to the objectivity of the test and to highlight the areas where students are conceptually weak" (Ahmad & Jamil, 2019. p. 90). This analysis is made up of three main activities: analyzing item difficulty levels, item discrimination power, and distractor effectiveness (Reynolds, Livingston, & Wilson, 2009). Reynolds, Livingston, and Wilson also stated that good validity and reliability of tests depend on the quality of the items on the test. For this reason, good quality items will create good quality tests.

However, most teachers are not that familiar with the skills of developing tests. Quansah, Amoako, and Ankomah (2019) found that teachers' ability to compose tests still lacked various aspects such as validity, reliability, and fairness after research in Kenya. They stated that the teachers lack training and knowledge about the subject. Unfortunately, not only some teachers, but some experts in evaluation also do not have the quality assurance and knowledge that is necessary to review more about this subject (Harris-Huemmert,

2011). Some teachers also seem to disregard the importance of evaluation and assessment. As mentioned by Styron & Styron (2012), most think that test-making training is the same as teaching methods and classroom management training, so they mostly do not go. There is also the case of item analysis in tests. As mentioned, the test item analysis is important, but, in most cases, item analysis which includes determination of item difficulty and discrimination, as well as distractor analysis are not done because it is time-consuming and demanding if done manually" (Tan, Cordova, Saligumba, & Segumpan, 2019, p. 63). Tests are the reflection of students' progress and development, so this problem needs to be so quickly.

The same problem has also been faced by a variety of teachers in Indonesia or locally. Jabbarifar (2009) wrote how learners' willingness to study and their enthusiasm can be heightened by a well-managed assessment and evaluation because they know how well they are doing in that class. Unfortunately, the case here in the locals is pretty much similar. Based on research by Masruroh (2014) at a school in Tulungagung, it was found that the test in that school was weak in validity (content and construct validity) and that it was too easy. Some teachers have admitted that there were no specific procedures and guidelines to help them in developing self-made tests. Anwar (2018) explained how there are teachers, especially in Indonesia, who developed test items that are either too easy or too difficult which makes the instrument of measurement underestimates or overestimates students' real performance. As a result, some students who had studied hard got results that did not reflect their skills, mostly because the tests were not designed according to the materials or key points that the teacher had taught. That is why teachers must pay careful attention to this matter and the problem of constructing test items.

A similar paradigm that teachers do not pay special attention to constructing quality test items was also found at a local private school in Manado. Based on a pre-interview with a concerned English teacher there, the test items development was still considered less important compared to other things, such as classroom management and methods of teaching. This was the reason why it was important to do research there to help the teacher be aware of the importance of constructing a good test and developing a good quality English test that the teacher can use. This study focused on the development of test items for Grade X at the school. The test items developed were part of a teacher-made test that was specifically created for the English 1st Mid-Term Test School Year 2019/2020. The test developed would later be analyzed to find the quality of the test items by using test item analysis on validity, reliability, level of item difficulty, item discrimination indices, and distractors effectiveness.

RESEARCH METHODS

This current study was a quantitative developmental research method involving statistical analyses. It sought to develop a test as a measurement instrument. Developmental research is "the systematic study of designing, developing and evaluating instructional programs, processes, and products that must meet the criteria of internal consistency and effectiveness" (Seels & Richey, 1994, p. 127). In other words, the test is a product developed specifically only for this study, and it must meet the criteria.

Population and Respondents

The population of this study was the tenth-grade students at a private senior high school in Manado. This study was specifically done among all the tenth-grade students who were studying at the school during the academic year of 2019/202. Grade X consisted of five classes. These were made up of Grade X MIPA 1 (27 students), Grade X MIPA 2 (26 students), Grade X MIPA 3 (27 students), Grade X IPS 1 (31 students), and Grade X IPS 2 (30 students). So, the total number of respondents for this study was 141 respondents. However, two students were not present during the day of the test, so the total number of respondents for

this study became 139 students. Then after checking the test, it was found that six students did not answer the items. Consequently, the test sheets of those who did not answer all the test items were excluded from this research. Finally, the respondents who participated in this research were 133 students.

Table 1. Respondents Demographic Data

No	Grade	Number of Respondents
1	X IPA 1	26
2	X IPA 2	26
3	X IPA 3	25
4	X IPS 1	29
5	X IPS 2	27
Total	5 Grades	133 Respondents

Instrument

The study used an English achievement test instrument that was developed for this study involving the English teacher at the senior high school. The test consisting of 50 multiple choice items was then tried out on the students during their English Mid Term Test period for Grade X of the academic year 2019/2020. But after the validity analysis, it became only 43 items that were qualified for the next step of analysis.

Test Development Procedure

The test development followed the following procedures:

1. Determining the learning indicators to test as a basic guide to develop the test.
2. Building the table of specifications or blueprints to help develop the test items easily.
3. Developing the test based on the blueprint, which was later sent to the concerned teacher via WhatsApp to be sorted and combined to make a full test.
4. Validating the finished test with the help of lecturers from a university.
5. Revising the test once more and later was tried out on the students during their Mid Term Test.
6. Receiving the results of the test from the teacher to be checked and analyzed.
7. Importing and analyzing the data using certain formulas by requesting help from a statistician.
8. Interpreting the results.

Data Analysis Techniques

The data gathered from the respondents was then analyzed and interpreted using statistical software. Data analysis and interpretation were mostly inferring, using tables, figures, and pictures to summarize the results, and discussing the results to answer the research questions (Creswell, 2012). The data analysis mostly used formulas that were created concerning the quality of the test.

To prove the validity, this study used Aiken's-V formula (Aiken, 1985) from the validation results that conducted by three expert raters, lecturers of the English department of the Faculty of Education at a private university. The lecturers evaluated the test items to prove whether the items tested were valid or not. The formula used was as follows:

$V = \Sigma s / [n(c-1)]$ (1)

$S = r - lo$

where
s= score given by the validators – lowest validity rate
r= score given by the validators (each item)
lo= lowest score given by the validators
n= number of validators
c= highest score given by the validators

After getting the results of each items' validity indices, those numbers was interpreted based on the scales provided by Retnawati (2016) (see Table 2).

Table 2. The Interpretation of Validity Analysis

Validity Index	Interpretation
≤ .4	Low validity
.4 – .8	Moderate validity
> .8	High validity

For measuring reliability, this study used Coefficient-Alpha with the following formula (Carr, 2012):

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum s_i^2}{s_x^2}\right) \quad (2)$$

where
k = the number of items on the test
s_i^2 = the population variance
$\sum s_i^2$ = the sum of all these item variances
s_x^2 = the population variance of the total score

If the results analyzed with the Coefficient Alpha formula show a score of .70 or higher than that, it has good reliability. Conversely, the test would not be reliable if the index is below .70 (Reynolds, Livingston, & Wilson, 2009). In other words, the test does not meet the requirement of reliability.

To calculate the difficulty levels of items, a formula proposed by Kubizyn and Borich (2013) was used. To know whether the items were too easy, average, or hard, the formula used was as stated below:

$$P = \frac{NP}{N} \quad (3)$$

where
P= level of difficulty
NP= the right response
N= the number of students

The results are interpreted based on the scales shown in the table below (Daryanto, 2018):

Table 3. The Interpretation of Item Difficulty Level

Level of Difficulty	Interpretation
.000 – .300	Difficult
.301 – .700	Moderate
.701 – 1.000	Easy

Before determining the lower-level and the higher-level students, the study sorted the scores of students from the highest to the lowest. Then the scores of students were

separated by taking 27% of the upper group and similarly 27% of the lower group (Kelley, 1939). The discrimination indices were then counted using the following formula:

$D = P_T - P_B$ (4)

where

D = Discrimination Power

P_T = proportion of the top group getting the answer correct

P_B = proportion of lesser mastery examinees getting the answer correct

(Reynolds, Livingston, & Wilson, 2009)

After finding the results using the formula, the score will later be interpreted following the discrimination criteria by Ebel (as cited in Sary, 2018) as seen in Table 4.

Table 4. The Interpretation of Item Discrimination Indices

Discrimination Criteria	Interpretation
.40 and above	Very good items; acceptable items
.30-.39	Reasonably good items but subject to improvement
.20-.29	Marginal items usually need and subject to improvement
Below .19	Poor items to be rejected or improved by revision

Distractors can be called good distractors if they function well. The indication that distractors function well, they must at least be chosen by around 5% of the whole students who go through that test (Sudijono, 2011). The formula that is used to find the effectiveness of the distractors is as follows (Arifin, 2012):

$$\frac{ED = P \times 100\%}{(N_1\text{-}C) / (n\text{-}1)} \quad (5)$$

where

ED = Effectiveness of distractors

P = Total of examinees who chose the distractors

N_1 = Total of examinees who joined the test

C = Total of examinees who correctly answer each item

N_2 = Total alternative answers

1 = Constant number

After measuring the distractors' effectiveness, the results were analyzed and interpreted based on the scales seen in Table 5 below (Arifin, 2012).

Table 5. The Interpretation of Distractors' Effectiveness

Effectiveness of Distractors	Interpretation
≥76%	Very good distractors
51% - 75%	Good distractors
26% - 50%	Mediocre distractors
0% - 25%	Poor distractors

RESULTS AND DISCUSSIONS

The results of content validity were gained by using Aiken's-V formula, where the total of items rater partake in this research was three validators with five levels of items rating. As suggested by Aiken (1985), the minimum acceptable score of the content validity of each item is .92. This research used the interpretation scales provided by Retnawati (2016) which categorize the numbers into three types of level: (1) high; (2) moderate; and (3) low. In this research, the acceptable or valid items are those which fell under the category of items with high V values.

Most of the items, as shown in Table 6, were found to have good validity. After being calculated and analyzed, 43 items were revealed to have high validity levels while seven items had moderate validity levels and must be removed because they did not reach the standard of V-Aiken validity which is at least .92 or more. The seven items were items 18, 19, 32, 33, 35, 38, and 4. These seven items were later removed and erased from the test and removed from the further test quality analysis. Related to this, Mutmaina (2017) in the content validity analysis similarly found three invalid items that they should be removed from the test. Alternatively, if the test developers want to use the items again, they should be improved to meet the validity criteria.

The Difficulty Level of Each Test Item

This part discusses the results of the level of difficulty of each test item. The level of difficulty is measured using the formula from Kubizyn and Borich (2016). The level of difficulty was then calculated and analyzed using the statistical tool. The results of each item after being counted in the statistical tool were later interpreted using the level of difficulty interpretation by Daryanto (2012). The interpretation was categorized into three levels: (1) easy; (2) moderate; (3) difficult. This interpretation would later determine whether the test items were difficult or easy.

Based on the results from Table 7, the test items were quite equal in these three levels. However, the easy test items were far too many compared to the moderate and difficult test items which mean that this test underestimated the students' ability. The test items that were categorized as easy test items were 31 items which included items 1, 2, 3, 4, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 20, 21, 22, 23, 24, 25, 27, 29, 31, 34, 41, 42, 43, 45, 47, 48, and 5. For the moderate level of difficulty, this test has 11 items included there. Those items were items 4, 5, 9, 17, 26, 28, 30, 36, 37, 44, 46, and 49. Finally, for the difficult category, there was only item 39. The low amount of difficult-level test items is good because that means that this test does not overestimate the students' ability which can result in not being able to answer each test item even when they have studied hard enough.

To prove the current study is supported by other studies, this study is compared with a study from Mahirah, Ahmad, and Sukirman (2016). The previous study found that 18 items were considered easy, 17 items were average or moderate, and five items were deemed difficult. This study and the previous one both found that there were more easy items than difficult ones. However, the previous study had a balanced item between easy and moderate ones, while the current study had a huge gap in the number of test items between easy and moderate levels. This means that the test developed for the current study must be revised and improved to increase the number of moderate items.

Table 6. Level of Content Validity

Item	Aiken's V	Interpretation
1	1.00	High
2	1.00	High
3	.92	High
4	1.00	High
5	.92	High
6	.92	High
7	.92	High
8	.92	High
9	.92	High
10	.92	High
11	1.00	High
12	1.00	High
13	1.00	High
14	1.00	High
15	1.00	High
16	1.00	High
17	1.00	High
18	.83	Moderate
19	.83	Moderate
20	.92	High
21	1.00	High
22	1.00	High
23	1.00	High
24	1.00	High
25	1.00	High
26	1.00	High
27	1.00	High
28	1.00	High
29	.92	High
30	.92	High
31	.92	High
32	.50	Moderate
33	.50	Moderate
34	.92	High
35	.83	Moderate
36	1.00	High
37	1.00	High
38	.67	Moderate
39	1.00	High
40	.75	Moderate
41	1.00	High
42	1.00	High
43	1.00	High
44	1.00	High
45	1.00	High
46	1.00	High
47	1.00	High
48	1.00	High
49	1.00	High
50	1.00	High

Table 7. Difficulty Levels of Each Test Item

Item	Level of Difficulty	Interpretation
1	.72	Easy
2	.83	Easy
3	.73	Easy
4	.71	Easy
5	.48	Moderate
6	.74	Easy
7	.99	Easy
8	.77	Easy
9	.70	Moderate
10	1.00	Easy
11	.97	Easy
12	.93	Easy
13	.90	Easy
14	.81	Easy
15	.84	Easy
16	.85	Easy
17	.55	Moderate
20	.88	Easy
21	.99	Easy
22	.84	Easy
23	.82	Easy
24	.93	Easy
25	.81	Easy
26	.69	Moderate
27	.81	Easy
28	.30	Moderate
29	.90	Easy
30	.62	Moderate
31	.90	Easy
34	.77	Moderate
36	.59	Moderate
37	.45	Moderate
39	.26	Difficult
41	.84	Easy
42	.89	Easy
43	.74	Easy
44	.42	Moderate
45	.78	Easy
46	.56	Moderate
47	.84	Easy
48	.81	Easy
49	.62	Moderate
50	.79	Easy

Discrimination Power Level of Each Test Item

The discrimination power level which determines whether the test items can differentiate the upper-level students and lower-level students was analyzed using a formula. Since all the test items were multiple-choice items, the formula used was the one proposed by Reynolds, Livingston, and Wilson (2009). The students' scores were first listed from the highest point until the lowest point and then divided into the upper class and the lower class by using 27% of each group (upper and lower). Then using the formula with the statistical tool, the data were analyzed and would be later interpreted using the interpretation by Ebel (as cited in Sary, 2018). This would then show whether the test items

could discriminate between the students who belong to the upper level and those who belong to the lower level.

Table 8. Discrimination Power Level of Each Test Item

Item	Discrimination Index	Interpretation
1	.72	Very Good
2	.50	Very Good
3	.67	Very Good
4	.39	Reasonably Good
5	.61	Very Good
6	.61	Very Good
7	.00	Poor
8	.61	Very Good
9	.67	Very Good
10	.00	Poor
11	.17	Poor
12	.28	Poor
13	.28	Poor
14	.33	Poor
15	.61	Very Good
16	.56	Very Good
17	.78	Very Good
20	.56	Very Good
21	.06	Poor
22	.67	Very Good
23	.61	Very Good
24	.22	Marginal
25	.67	Very Good
26	.78	Very Good
27	.56	Very Good
28	.78	Very Good
29	.17	Poor
30	.56	Very Good
31	.50	Very Good
34	.50	Very Good
36	.44	Very Good
37	.39	Reasonably Good
39	.33	Reasonably Good
41	.50	Very Good
42	.28	Marginal
43	.67	Very Good
44	.78	Very Good
45	.50	Very Good
46	.72	Very Good
47	.44	Very Good
48	.67	Very Good
49	.61	Very Good
50	.56	Very Good

The results from Table 8 above showed that the overall discrimination power of each test item was very good with only some items being poor. Thirty items were considered "very good" in terms of discrimination power. Next, there were three items whose indices were categorized as "reasonably good" which need a bit of further improvement. For marginal items, which needed to be improved, it contained only one item and for the poor items, it had eight items. The item discrimination power indices became "poor" for the eight items, mostly not because it cannot discriminate, but mostly because most students were able to answer the items correctly in both the upper and lower class. This result correlated

with a study by Shomami (2014) which found seven poor items and one negative item which must be erased and 16 good items for discrimination indices. The previous study had quite good test items in terms of discrimination power. In comparison with the previous study, the test had very good item discrimination indices for it had many items which were considered very good and had only a small number of poor items. These items also did not have any negative index, meaning that the test was very good in its ability to separate the upper group from the lower group.

Effectiveness of the Distractors of the Multiple-Choice Items

This part answers the question of whether the distractors or alternatives of each test item were effective in confusing the students who did not master the materials they were studying. The formula used to determine the effectiveness of each distractor is by Arifin (2012) who computed the percentage of each distractor in correlation with the answer key and the number of alternatives given. The interpretation is categorized into five which are: (1) very good, (2) good, (3) mediocre, (4) poor, and (5) very poor. This interpretation is also by Arifin (2012). It is to be noted that each distractor cannot possibly become a very good distractor given the nature of the students' condition in answering each test item.

Overall, the results of analyses of the distractors per item showed that the distractors overall were quite good. Even though the alternatives were quite a lot, and students might tend to pick the dominant one among other possible distractors, the distractor distributions were quite decent. The distractors categorized as very good were 42 distractors, while the "good" ones were 22 distractors. The mediocre ones consisted of 38 distractors, while the poor and very poor distractors consisted of 48 and 18 distractors respectively. The very poor ones were the result of the students eliminating each alternative and finally ending up choosing the same distractors. There was one test item in which the distractors did not function at all, mainly because all the students chose the answer correctly. It correlates with a study by Mutmaina (2017) which also found the effectiveness of the test to be quite good. Among 76 distractors, the distractors in two items were considered very good, three items in which their distractors were considered good, and seven items got mediocre distractors. There were six items considered poor and not very poor distractors. This is understandable in comparison with the current study. The previous study only analyzed 19 items in comparison to the 37 items for the current study. Overall, the test distractors were quite good because 64 distractors can be said to be good. The 38 distractors were moderate in terms of distracting the students, and 66 distractors were the "poor" ones. However, the poor ones still need improvement so the distractors can function better.

Reliability of the Whole Test

To measure the reliability level of the whole test, this research used the Cronbach Alpha formula. Using the statistical tool, the research did not calculate the Alpha coefficient manually but analyzed the data immediately using the software built-in device. As mentioned in Reynolds, Livingston, and Wilson (2009), a test can be considered reliable if its reliability index is equal to or more than .7. Based on the Cronbach Alpha analyses, the test developed was reliable because the coefficient alpha was .89 which was higher than .70. It is proven by Kusuma (2010) who mentioned that test reliability will be good if the items are good. It is also supported by another research done by Opara and Magnus-Arewa (2017) who found that their test items developed were also reliable with a .73 level of reliability. To compare, the test results for the current study were reliable because the item analysis results were also quite good.

CONCLUSIONS

After analyzing the content validity of the items and the reliability of the test, the test items went through an item analysis process. The summary of the analysis results is shown in Table 1. It summarizes the interpretations of each item's difficulty level, discrimination index, and distractor effectiveness.

Table 1. Description of the Item Analysis

Item	Difficulty	Discrimination	Description of the Distractors' Effectiveness
1	Easy	Very Good	Two are effective; two need improvement
2	Easy	Very Good	Two are effective; two need improvement
3	Easy	Very Good	All distractors need improvement
4	Easy	Reasonably Good	One is effective; three need improvement
5	Moderate	Very Good	All distractors need improvement
6	Easy	Very Good	Two are effective; two need improvement
7	Easy	Poor	All distractors need improvement
8	Easy	Very Good	One is effective; three need improvement
9	Moderate	Very Good	Two are effective; two need improvement
10	Easy	Poor	All distractors are not functional
11	Easy	Poor	All distractors need improvement
12	Easy	Poor	All distractors need improvement
13	Easy	Poor	Two are effective; two need improvement
14	Easy	Poor	One is effective; three need improvement
15	Easy	Very Good	Two are effective; two need improvement
16	Easy	Very Good	Two are effective; two need improvement
17	Moderate	Very Good	Two are effective; two need improvement
20	Easy	Very Good	One is effective; three need improvement
21	Easy	Poor	All distractors need improvement
22	Easy	Very Good	Two are effective; two need improvement
23	Easy	Very Good	Two are effective; two need improvement
24	Easy	Marginal	Two are effective; two need improvement
25	Easy	Very Good	Two are effective; two need improvement
26	Moderate	Very Good	Two are effective; two need improvement
27	Easy	Very Good	One is effective; three need improvement
28	Moderate	Very Good	All distractors need improvement
29	Easy	Poor	Three are effective; one needs improvement
30	Moderate	Very Good	All distractors need improvement
31	Easy	Very Good	Two are effective; two need improvement
34	Moderate	Very Good	Two are effective; two need improvement
36	Moderate	Very Good	Three are effective; one needs improvement
37	Moderate	Reasonably Good	One is effective; three need improvement
39	Difficult	Reasonably Good	Three are effective; one needs improvement
41	Easy	Very Good	All distractors are effective
42	Easy	Marginal	One is effective; three need improvement
43	Easy	Very Good	All distractors need improvement
44	Moderate	Very Good	Three are effective; one needs improvement
45	Easy	Very Good	Two are effective; two need improvement
46	Moderate	Very Good	One is effective; three need improvement
47	Easy	Very Good	All distractors are effective
48	Easy	Very Good	Two are effective; two need improvement
49	Moderate	Very Good	One is effective; three need improvement
50	Easy	Very Good	One is effective; three need improvement

Based on the results in the table, there are only 11 items that can be considered qualified in terms of item difficulty levels and discrimination power out of 50 test items. However, these items are not fully qualified because each of the eleven items has one, two, or three non-functioning distractors or distractors which are ineffective in distracting the students. It can be because all the students studied hard for the test, or it can be some other factors that determine the students' ability to perform well in tests. However, that is why it is important to do item analysis so that the test questions can be improved even better.

RECOMMENDATIONS

Teachers can directly use the items whose difficulty levels and discrimination power are good in testing with a little improvement on their nonfunctioning distractors. If teachers want to use unqualified items, those whose difficulty levels and discrimination power are

below the standard, they need to improve these items. They should be improved in terms of difficulty levels, discrimination power, and nonfunctioning distractors. When they were improved, they should be tried out again. Then, the tryout results should be analyzed again in terms of validity, reliability, difficulty level, discrimination power, and distractor effectiveness to ensure the items' quality.

REFERENCES

Ahmad, S., & Jamil, S. (2019). Analysis of test items used in an achievement test in Physics at secondary level. *Journal of Education and Practice, 10*(10), 90-96. doi: 10.7176/JEP

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *SAGEPUB Social Science Journal, 45*, 131-142. doi: 10.1177/ 0013164485451012.

Anwar, M. (2018). *Menjadi guru profesional.* Jakarta, Indonesia: Prenadamedia Group.

Arifin, Z. (2012). *Evaluasi pembelajaran*. Bandung, Indonesia: Remaja Rosdakarya.

Baranovskaya, T., & Shaforostova, V. (2017). Assessment and evaluation techniques. *Journal of Language and Education*, 3(2), 30-38. doi: 10.17323/2411-7390-2017-3-2-30-38.

Carr, N. T. (2011). *Designing and analyzing language tests.* Oxford, England: Oxford University Press.

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Boston, MA: Pearson.

Daryanto. (2012). *Penelitian tindakan kelas dan penelitian tindakan sekolah: Beserta contoh-contohnya*. Yogyakarta, Indonesia: Gava Media.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Grayson, T. E. (2012). Program evaluation in higher education. In C. Secolsky, & D. B. Denison (Eds.), *Handbook on: Measurement, assessment, and evaluation in higher education* (pp. 459-472). New York, NY: Taylor & Francis.

Harris-Huemmert, S. (2011). *Evaluating evaluators: An evaluation of education in Germany.* Wizbaden, Germany: Springer Fachmedien.

Hussain, S., & Sajid, S. (2015). Test construction and evaluation: A brief review. *Indian Journal of Applied Research, 5*(6), 725-729. Retrieved from: https://www.researchgate.net/publication/314894221_Test_Construction_and_Evaluation_A_Brief_Review/link/58c70b99aca27232ac824cc7/download

Jabbarifar, T. (2009, November 16-18). The importance of classroom assessment and evaluation in educational system. Paper presented at 2nd International Conference of Teaching and Learning, Malaysia, 2009. Kuching, Malaysia: INTI University College. Retrieved from: https://pdfs.semanticscholar.org/db8c/4d3e5e56aa80c220e17eeac25183acaaa43d.pdf

Karkal, Y. R., & Kundapur, G. S. (2016). Item analysis of multiple choice questions of undergraduate pharmacology examinations in an International Medical School in India. *Journal of Dr. NTR University of Health Sciences, 5*(3), 183-186. Retrieved from: https://www.researchgate.net/publication/309000992_Item_analysis_of_multiple_choice_questions_of_undergraduate_pharmacology_examinations_in_an_International_Medical_School_in_India

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology 30*(1), 17-24. Retrieved from: https://psycnet.apa.org/buy/1939-03313-001

Kubiszyn, T., & Borich, G. (2013). *Educational testing and measurement: Classroom application and practice* (10th ed.). Danvers, MA: John Wiley & Sons.

Kusuma, M. (2010). *Evaluasi pendidikan.* Jakarta, Indonesia: Multi Kreasi Satudelapan.

Kusumawati, M., & Hadi, S. (2018). An analysis of multiple choice questions (MCQs): Item and test statistics from mathematics assessments in senior high school. *Research*

and Evaluation in Education, 4(1), 70-78. Retrieved from: https://journal.uny.ac.id/index.php/reid/article/view/20202/11492

Mahirah, R., Ahmad, D., & Sukirman. (2016). Designing multiple choice test of vocabulary for the first semester students at English Education Department of Alauddin State Islamic University of Makassar. *Eternal (English, Teaching, Learning, and Research Journal)*, *2*(2). https://doi.org/10.24252/Eternal.V22.2016.A9

Masruroh, H. Z. (2014). An item analysis of English summative test for second grade students of MAN Tulungagung 1 in Academic Year 2013/2014 (Unpublished thesis). State Islamic Institute, Tulungagung, Indonesia. Retrieved from: http//repo.iain-tulungagung.ac.id/707/.

Munadliroh, S. (2015). Items analysis on the score of the English summative test (Unpublished thesis). State Institute for Islamic Studies, Salatiga, Indonesia. Retrieved from: https://www.pdfdrive.com/items-analysis-on-the-score-of-the-english-summative-test-e53206411.html

Mutmaina, D. (2017). Pengembangan instrumen test diagnostik pilihan ganda dua tingkat untuk mengidentifikasi pemahaman konsep Matematika Wajib siswa MAN 1 Makassar (Unpublished thesis). Universitas Islam Negeri Alauddin Makassar, Makassar, Indonesia. Retrieved from: http://repositori.uin-alauddin.ac.id/7818/

Opara, I. M., & Magnus-Arewa, E. A. (2017). Development and validation of mathematics achievement test for primary school pupils. *British Journal of Education, 5*(7), 45-47. Retrieved from: http://www.eajournals.org/wp-content/uploads/Development-and-Validation-of-Mathematics-Achievement-Test-for-Primary-School-Pupils.pdf

Quansah, F., Amoako, I., & Ankomah, F. (2019). Teachers' Test Construction Skills in Senior High Schools in Ghana: Document Analysis. *International Journal of Assessment Tools in Education, 6*(1), 1-8. doi: 10.21449/ijate.481164

Retnawati, H. (2016). *Analisis kuantitatif instrument penelitian: Panduan peneliti, mahasiswa, dan psikometrian*. Yogyakarta, Indonesia: Parama Publishing.

Reynolds, C. R., Livingston, R. B., & Wilson, V. (2009). *Measurement and assessment in education: International edition* (2nd ed.). London, England: Pearson.

Sari, A. Y. (2017). Item analysis of English Mid-Term test items for the second semester of the seventh grade students of SMP Negeri 2 Wonosari in the 2015/2016 Academic Year (Unpublished thesis). State Islamic Institute of Surakarta, Surakarta, Indonesia. Retrieved from: http://eprints.iain-surakarta.ac.id/662/1/29.%20Anis%20 Yunita%20Sari.pdf

Seels, B. B., & Richey, R. C. (1994). Instructional technology: The definition and domains of the field. Washington, DC: Association for Educational Communications and Technology.

Shomami, A. (2014). An item analysis of English summative test: An Analysis Study in the second grade of SMA Negeri 6 Depok in the 2013/2014 academic year (Unpublished thesis). Universitas Islam Negeri Syarif Hidayatullah Jakarta, Jakarta, Indonesia.

Styron, J. L., & Styron, R. A. (2012). Teaching to the test: A controversial issue in quantitative measurement. *Systemics, Cybernetics, and Informatics, 10*(5). 22-25. Retrieved from: http://www.iiisci.org/Journal/CV$/sci/pdfs/ HEA561DK.pdf

Sudijono, A. (2011). *Pengantar evaluasi pendidikan*. Jakarta, Indonesia: RajaGrafindo Persada.

Tan, D. A., Cordova, C. C., Saligumba, I. P. B., & Segumpan, L. L. B. (2019). Development of valid and reliable teacher-made tests for grade 10 Mathematics. *International Journal of English and Education, 8*(1), 62-83. Retrieved from: https://www.researchgate.net/publication/328556552_ Development_of_Valid_and_Reliable_Teacher-Made_Tests_for_Grade_10_ Mathematics